



February 8, 2000

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #O- 5

MEMORANDUM FOR Michael J. Longini
Chief, Decennial Systems and Contracts Management Office

From: Howard Hogan *Howard Hogan*
Chief, Decennial Statistical Studies Division

Prepared By: Kimball Jonas *KJ*
Data Processing Techniques and Requirements Branch

Subject: Transcription of Group Quarters (GQ) ID Numbers on GQ
Enumeration Forms

I. INTRODUCTION

The forms used for GQ enumeration do not go to the field with Master Address File (MAF) IDs on them. The means for linking the form to its geography is the fourteen-digit GQ ID, which is hand-copied by the enumerator from a master form onto each enumeration form. The potential for error in hand-transcribing such a long number is of concern. With an expected processing volume of 9.7 million such forms in Census 2000, there is a realistic possibility of several hundred thousand forms with invalid GQ IDs. At present, the plan is that forms captured with erroneous GQ IDs will not be processed; instead, persons will be imputed to substitute for the lost data.

Currently, the only measure proposed to reduce the number of such transcription errors that reach data capture is a series of visual checks by Census 2000 field workers and clerical staff, comparing the GQ IDs on the enumeration forms against the eye-readable GQ ID on the computer-printed D-352 master form.

The handwritten GQ ID field is interpreted by Optical Character Recognition (OCR) software. At the interpretation stage of data capture, a check digit test will be run on the GQ IDs, identifying which ones are incorrect. However, as the process is currently set up, the master form with the correct GQ ID will not be available to correct transcription errors when the failures are found not to be in the OCR interpretation.

The purpose of this memorandum is to examine the process by which GQ IDs are checked, to examine and evaluate possible improvements to this process, and to make a recommendation with the goal of minimizing the number of GQ enumeration forms that go unprocessed due to mis-transcribed GQ IDs.

Refer any comments or questions to Kim Jonas of my staff, at (301) 457-4197.

II. GQ FORMS: ENUMERATION, DATA CAPTURE, AND PROCESSING

A D-352 master form is generated by the operational control system, OCS 2000, for each GQ to be enumerated. It has the correct GQ ID on it, and it accompanies the enumeration forms from the GQ, from enumeration through the check-in process at the National Processing Center (NPC).

The enumerator must copy the GQ ID, a fourteen-digit number, from the D-352 to each GQ enumeration form.¹ This is true for all of the Individual Census Questionnaires (ICQs), Individual Census Reports (ICRs), and Military Census Reports (MCRs), an expected 9.7 million forms in all. A detailed list is appended as Attachment A.

The GQ IDs are transcribed by the enumerators prior to enumeration. Afterwards, the forms go through clerical checks, most of which include visual checks comparing the GQ ID on the form to the GQ ID on the D-352. In the case of ICRs, (i) the crew leader, and (ii) the LCO clerical staff, visually check the GQ ID on each form as part of their overall checks of the enumerator's work, before the LCO checks out the GQ.

Depending on the size of the GQ, the ICRs and D-352s from several GQs may be put in one box, which is checked out and shipped to the Jeffersonville DCC. The DCC checks the box in, then checks in each GQ represented in the box, by keying in the GQ ID on the D-352. The process of transcription and clerical checks applicable to the ICRs is described in more detail in Attachment B; the process for the ICQs and MCRs is similar to that for the ICRs.

A computer analysis of the MCRs and short form ICQs and ICRs from the 1998 Dress Rehearsal (DR) shows GQ ID error rates of between 0.6% and 2.3% (excluding blank GQ ID fields), varying with form type².

After the check-in at the Jeffersonville Data Capture Center (DCC), the ICQs go to the Non-DCC³ for keying, while the ICRs and MCRs stay at the DCC for capture by Lockheed Martin's DCS2000 data capture system. The ICQs will be checked in at the Non-DCC, both by GQ and by individual form. The GQ ID from each ICQ will be keyed and checked by computer against the master list of valid GQ IDs. If it fails this

¹ The GQ ID is formatted with dashes (e.g. 01234-56789-0123) on the D-352, but is written without breaks (e.g. 01234567890123) on the GQ enumeration form.

² DR normalized files. Error rates were 0.7%, 0.6%, and 2.3% for MCRs, short form ICRs, and short form ICQs, respectively. Long form ICQs and ICRs were not analyzed. A more detailed breakdown is attached as Attachment D.

³ The Non-DCC is that portion of the NPC outside the DCC.

check, troubleshooters at the Non-DCC will attempt to match ICQs to a GQ. They will not use the GQ ID on the D-352 to correct the GQ ID on the ICQ, but will instead try to find the best match from the SP/GQ master list, though the D-352 is readily available.

If the GQ ID on an ICR or MCR fails the check digit test because the GQ ID was incorrectly written down, there is no plan to try to match the record with a GQ. Instead, imputed persons will be added to those GQs that have lost population during data capture.

III. RECOMMENDATIONS

In order to best attempt to reunite a GQ enumeration form with an invalid GQ ID with its original GQ, as much information as is available regarding the source of that GQ form should be used. If the D-352 is available when the erroneous GQ ID is discovered, then the GQ ID on the D-352 should be used in correcting the GQ ID on the GQ form. If the D-352 is unavailable, there is another context-utilizing approach that can be applied.

A. Fixing the ICQs: Correction of GQ ID at Non-DCC Check-In

At Non-DCC check-in, the GQ ID is keyed in and electronically compared with the list of valid GQ IDs. At this point, the D-352 is still available for comparison. An invalid GQ ID can be corrected here by (i) having the check-in keyer write the GQ ID from the D-352 above the incorrect GQ ID, and (ii) having the troubleshooting unit⁴ correct the invalid GQ ID on the ICQ to the valid GQ ID on the D-352 where appropriate⁵.

B. Fixing the ICRs and MCRs: Within-Batch GQ ID Check at DSCMO

For ICRs and MCRs, there appears to be no easy way to compare the GQ ID on the form with the GQ ID on the corresponding D-352. But it is still possible to preserve some context for a given ICR or MCR by attempting to match invalid GQ IDs on these forms to valid GQ IDs in the same batch.

Forms from a GQ will generally go to the scanner together, so it is likely that the GQ ID of the ICR or MCR's originating GQ is on that list. Since there will be perhaps 300,000 valid GQ IDs altogether, but fewer than 200 valid GQ IDs in a

⁴ The Address Register Review and Feedback (ARRF) Unit.

⁵ DSSD has no position at this time regarding how close an invalid GQ ID number should be to the correct GQ ID number of the GQ a form is apparently from, in order to be corrected to the valid GQ ID. At a minimum, though, invalid GQ IDs that differ from the D-352's GQ ID by one digit or by a transposition of consecutive digits should be corrected to the GQ ID on the D-352.

batch⁶, it makes sense to compare an invalid GQ ID on an enumeration form against the much shorter list rather than the longer one, especially since one would expect to find the correct GQ ID in the shorter list.

IV. CONCLUSION

The problem of incorrectly transcribed GQ IDs will be ameliorated somewhat by the visual checks described earlier, but it is sensible to try to remedy the problem after the GQ IDs have been checked for validity by computer software.

We have outlined a solution that should reunite the vast majority of such forms with their proper GQs in a simple and efficient manner. We recommend its adoption.

cc:

DSSD Census 2000 Procedures and Operations Memorandum Series

L.Affleck	(DSCMO)	J. Lothar	(GEO-ALRB)
D. Burkhead		E. Jaramillo	(CLO)
J. Dickens		S. Rolark	
G. McLaughlin		E. Byerly	(POP)
D. Stoudt		A. Smith	
J. Stovall		D. Smith	
J. Benetti	(FLD)	L. Kehm	
S. Schoch		L. Moyer	(DMD)
M. Weiler		D. Dalzell	(HHES)
J. Dawson	(TMO)	T. McNally	(DSSD)
K. Wyatt			
C. Falkenstein	(GEO-MAF))		
J. McKay			

⁶ Frequently much fewer: some batches may contain just one GQ, hence just one GQ ID.

GQ Enumeration Forms With Hand-Transcribed GQ IDs¹

Form Name	Form Number	Short/Long	Language	Data Capture Location	Data Capture Method	Expected Processing Volume
ICQ	D-15A	Short	English	Non-DCC ²	Keying ³	1,000,000
	D-15B	Long	English	Non-DCC	Keying	200,000
ICR	D-20A	Short	English	DCC	Scanning ⁴	5,700,000
	D-20B	Long	English	DCC	Scanning	1,000,000
	D-20A(S)	Short	Spanish	DCC	Scanning	700,000
	D-20B(S)	Long	Spanish	DCC	Scanning	50,000
MCR	D-21	Long ⁵	English	DCC	Scanning	1,000,000
Puerto Rico ⁶	various			Non-DCC	Keying	50,000

¹Source: "Forms Information (by Type of Operation, Data Capture Location, Bar Code, Etc.)" (table), prepared by Fabian Sanchez, DSCMO-DCPO-Lanham, October 4, 1999.

²"Non-DCC" is a designation for the NPC operations outside of the Jeffersonville DCC.

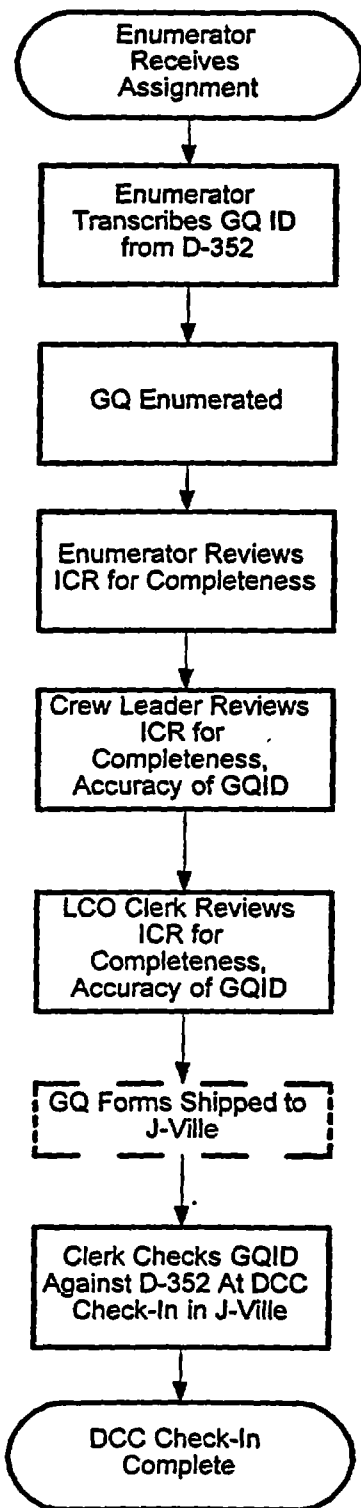
³Specifically, keying from the paper form, or KFP.

⁴The forms are scanned, and the images saved on disk. The DCC server's Optical Character Recognition/Optical Mark Recognition (OCR/OMR) software then reads the image in each field. If the software has sufficiently high confidence that the field has been captured correctly, the result is saved; if not, then the image of that field is sent electronically to Key From Image (KFI) keyers.

⁵The MCR is a hybrid: it's a long form, but respondents whose SSN (last 4 digits) is less than 8333 fill out only the 100% questions.

⁶This line represents the nine kinds of GQ forms used in Puerto Rico: short and long ICQs and ICRs in English and Spanish, and MCRs.

ICRs: Transcription and Clerical Checking of GQ ID



The enumerator fills out a GQ Listing Sheet, enters the GQ ID and a person number on each ICR, filling out the ICR envelope, puts the ICR and a Privacy Act notice inside.

The enumerator isn't expected to check the GQ ID's accuracy.

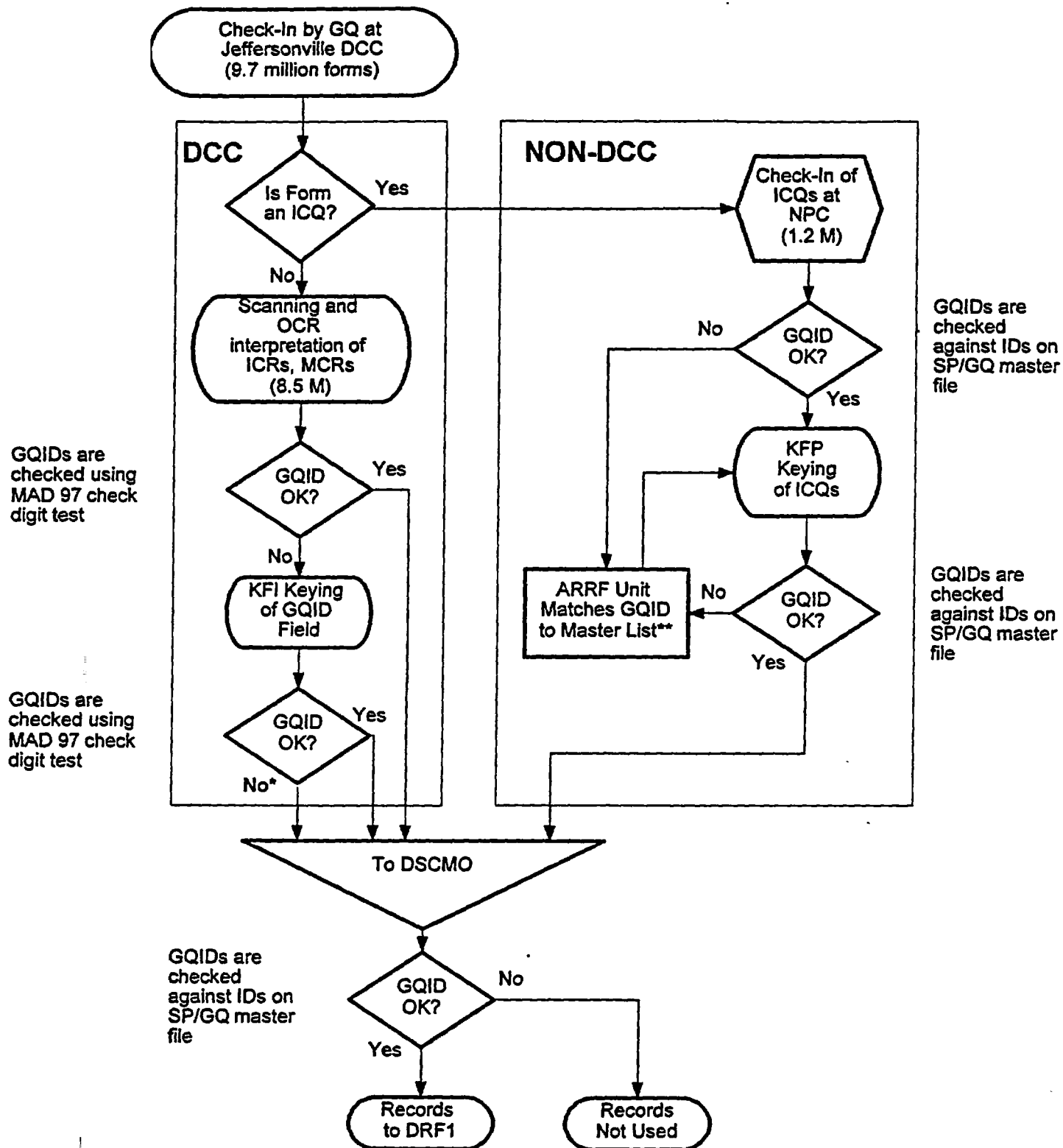
The Crew Leader checks the GQ Listing Sheet, ascertains that the ICRs correspond with the Listing Sheet, checks each ICR for completeness (including skip pattern on long forms), checks each ICR's GQID against the D-352, checks Enumerator Questionnaires, and reviews the enumerator's Daily Pay and Work Record.

The LCO review is similar to the Crew Leader's review. After passing the LCO review, the GQ is checked out of the LCO. The GQ pop is entered into OCS 2000 from the D-352 at LCO check-in. If LCO review corrects the pop count, the D-352 is corrected, and the new number is input into OCS 2000 at LCO check-out.

Each box containing GQ forms is checked in, then clerks check in the individual GQs represented in the box, counting the forms from each GQ to obtain a check-in count, and verifying the GQ IDs on each form against the D-352. Both the DCC check-in count and LCO check-out count are stored in OCS 2000; if they differ, discrepancies between the two are not reviewed at this stage.

GQ Enumeration Forms: Check-In and Capture at NPC

Attachment C



* ICRs and MCRs following this path will have failed the MAD 97 check digit test at OCR, been keyed in, re-tested, and failed again. These forms will go to DSCMO with an invalid GQID.

** ICQs that go to the ARRF Unit will be matched to a GQ on the basis of ID and address (if any). The GQ ID on the D-352 master form will not be taken into account.

GQID Errors in 1998 Dress Rehearsal Normalized Files

Record Type	ICQ short	ICR short	MCR
GQ Persons	975	28944	6825
ID with Blank	8	135	140
Solo ID, PN not 1	20	117	37
Fail MAD97 CD test	2	48	10
% ID with Blank	0.82	0.47	2.05
% Solo ID, PN not 1	2.05	0.40	0.54
% Fail MAD97 CD test	0.21	0.17	0.15
Total % Error	3.08	1.04	2.74
Total % Error, excluding IDs with blanks ¹	2.26	0.57	0.69

¹ Since IDs with blanks include a substantial number of blank IDs, many or most GQ ID errors in this category are not transcription errors.